Monolingual Post-Editing by a Domain Expert is Highly Effective for Translation Triage

Lane Schwartz lanes@illinois.edu
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana IL, USA

Abstract

Various small-scale pilot studies have found that for at least some documents, monolingual target language speakers may be able to successfully post-edit machine translations. We begin by analyzing previously published post-editing data to ascertain the effect, if any, of original source language on post-editing quality. Schwartz et al. (2014) hypothesized that post-editing success may be more pronounced when the monolingual post-editors are experts in the domain of the translated documents. This work tests that hypothesis by asking a domain expert to post-edit machine translations of a French scientific article (Besacier, 2014) into English. We find that the monolingual domain expert post-editor was able to successfully post-edit 86.7% of the sentences without requesting assistance from a bilingual post-editor. We evaluate the post-edited sentences according to a bilingual adequacy metric, and find that 96.5% of those sentences post-edited by only a monolingual post-editor are judged to be completely correct. These results confirm that a monolingual domain expert can successfully triage the post-editing effort, substantially reducing the workload on the bilingual post-editor by only sending the most challenging sentences to the bilingual post-editor.

1 Introduction

Post-editing is the process whereby a human user corrects the output of a machine translation system. The use of basic post-editing tools by bilingual human translators has been shown to yield substantial increases in terms of productivity (Plitt and Masselot, 2010) as well as improvements in translation quality (Green et al., 2013) when compared to bilingual human translators working without assistance from machine translation and post-editing tools. More sophisticated interactive interfaces (Langlais et al., 2000; Barrachina et al., 2009; Koehn, 2009b; Denkowski and Lavie, 2012) may also provide benefit (Koehn, 2009a).

Small-scale studies have suggested that monolingual human post-editors, working without knowledge of the source language, can also improve the quality of machine translation output (Callison-Burch, 2005; Koehn, 2010; Mitchell et al., 2013), especially if well-designed tools provide automated linguistic analysis of source sentences (Albrecht et al., 2009). Schwartz et al. (2014) confirmed this result with eight monolingual post-editors on a larger 3000 sentence test corpus.

Using a bilingual judge, we evaluate the post-edited test English sentences using the 10-point adequacy metric (see Table 5) of Albrecht et al. (2009). The results of our evaluation indicate that over 95% of post-edited sentences are completely correct translations that adequately convey the meaning of the respective French source sentence. Our bilingual judge estimated that approximately 15 minutes of total effort would be required for a bilingual French-English speaker to correct the remaining 5% of post-edited sentences.

2 Effects of Original Source Language on Post-Editing Quality

In discussing post-editing, there may be cases where shared task evaluation data may have an unintended effect on post-editing quality. When a shared task test set for a particular language pair (for example, from Russian into English) is created, some portion of that shared test set may have originally been written in the shared task target language, and then professionally translated into the shared task source language. By examining the data from the 2014 Workshop on Statistical Machine Translation, we have confirmed that this is indeed the case for (at least) the Russian-English shared task.

Schwartz et al. (2014) performed a post-editing experiment as part of the WMT 2014 Russian-English shared task. The post-editors in that study anecdotally reported an effect on post-editing difficulty based on original source language: Schwartz et al. noted:

> Interestingly, several post-editors self-reported that they could tell which documents were originally written in English and were subsequently translated into Russian, and which were originally written in Russian, based on observations that sentences from the latter were substantially more difficult to post-edit. Once per-document source language data is released by WMT14 organizers, we intend to examine translation quality on a per-document basis and test whether post-editors did indeed perform worse on documents which originated in Russian.

This effect, if it does indeed exist, could mean that positive post-editing results such as those reported by Schwartz et al. (2014) may be artificially high, due to the presence of sentences in the test set which were originally written in English. Such sentences may have maintained the original English word order even after translation through Russian, and so may have been easier to translate than sentences originally authored in Russian, which might be expected to be more difficult due to more idiomatic Russian word order.

Before exploring our own post-editing study in Section 3, we therefore find it useful to conduct some further data analysis on previously released data to attempt to ascertain what effect, if any, the original source language may play in post-editing quality. After the workshop, the WMT 2014 organizers released information regarding the original source language of each sentence in the shared task test sets. In addition, as part of their WMT 2014 submission, Schwartz et al. (2014) made available the post-edited translations from their Russian-English submission, along with the results of their manual evaluation.

Schwartz et al. (2014) report that their machine translations were post-edited by a group of eight individuals. We divide their post-edited translations by original source language and by post-editor, along with the binary adequacy judgements reported for each post-edited translation. Table 1 presents the results of this data collation. For each monolingual post-editor, the percentage of sentences judged to be correct according to a monolingual human judge are broken down according to the language in which test documents were originally authored. For 7 out of 8 post-editors, we observe worse translation quality for sentences originally authored in Russian when compared to sentences originally authored in English. The overall percentage of sentences judged to be correct, taken across all post-editors, is 14 percentage points lower for sentences originally authored in Russian (57% correct) when compared to sentences originally authored in English (71% correct). Interestingly, we see no coherent effect when quality is measured using BLEU (see Table 2); for some post-editors, BLEU scores are higher (more positive) for sentences originally authored in English, but for most post-editors, BLEU scores for some post-editors are higher (in some cases by more than 5 BLEU points) for sentences originally authored in Russian.

These partially contradictory results could be an artifact of metrics, or indicative of other factors at play. In Section 3, we examine one factor that may play a more important role in

	Post-Editor								
	1	2	3	4	5	6	7	8	All
en % correct	78%	67%	78%	62%	67%	48%	64%	72%	71%
ru % correct	65%	69%	52%	51%	63%	40%	60%	43%	57%

Table 1: For each monolingual post-editor in Schwartz et al. (2014), the percentage of sentences judged to be correct according to a monolingual human judge, broken down according to the language in which test documents were originally authored.

	Post-Editor								
	1	2	3	4	5	6	7	8	All
English	27.97	21.08	25.20	28.16	27.94	21.22	23.34	24.10	25.56
Russian	27.38	26.82	24.21	27.18	28.98	22.73	28.92	26.03	26.62
difference	0.59	-5.74	0.99	0.98	-1.04	-1.51	-5.58	-1.93	-1.06

Table 2: Analysis of post-edited translation data from Schwartz et al. (2014), showing case-sensitive BLEU scores per post-editor broken down according to the language in which test documents were originally authored.

predicting post-editing quality: domain expertise.

3 Monolingual Post-editing by a Domain Expert

It has been proposed (Schwartz et al., 2014) that post-editing machine translations may be more successful when the post-editor is highly familiar with the subject matter being translated. In this section we test that hypothesis by asking a domain expert to post-edit machine translations of a French scientific article (Besacier, 2014) into English.

We begin by copying the headers, content sentences, and other text comprising Besacier (2014) from the original PDF document into plain text format (UTF-8 encoding), dividing the text into 241 distinct segments.[1] To better facilitate machine translation, each segment was placed on its own line.

The plain text of the French source document was translated using Google Translate (Google, 2014), Systran Server 7.4.2 (Systran, 2010), and Moses (Koehn et al., 2007). Google Translate is a proprietary online statistical translation system that makes use of phrase-based translation methods. Systran Server is a proprietary translation system that is primarily rule-based, although recent versions allow for hybrid rule-based/statistical functionality; we did not make use of hybrid functionality in this experiment. Moses is the de-facto standard open source phrase-based statistical machine translation system. In our experiments, Moses was trained and tuned on French-English data from IWSLT 2013, following the procedures described in Kazi et al. (2013).

The monolingual post-editor is a native speaker of English with no training or experience in French with domain expertise in the scientific article being post-edited. For each sentence, the monolingual post-editor was presented with the machine translation results produced by the three aforementioned machine translation systems. The post-editor was free to choose any of the three MT output segments as the starting point for post-editing, and was free to incorporate portions of any or all of the three MT output segments into the final post-edited result. No

[1] While most of the segments are sentences, some segments are section headers, table elements, footnotes, etc. Throughout we will use the terms segment and sentence interchangeably.

Confident	The monolingual post-editor is confident that the post-edited translation conveys the meaning of the French sentence
Verify	The monolingual post-editor believes that the post-edited translation conveys the meaning of the French sentence, but the translation should be verified by a bilingual post-editor
Partially unsure	The monolingual post-editor is not confident that a specific portion of the post-edited translation is correct; that section should be handled by a bilingual post-editor
Completely unsure	The entire sentence should be handled by a bilingual post-editor

Table 3: Confidence guidelines for monolingual post-editors.

	Post-Editor Confidence			
	Completely unsure	Partially unsure	Verify	Confident
# sentences	8	13	11	209
% sentences	3.3%	5.4%	4.6%	86.7%

Table 4: Post-editor confidence in the adequacy of post-edited translations. Confidence labels are defined in Table 3.

interactive post-editing software was provided to the post-editor; for each sentence, the post-editor was presented with the three MT output segments, and was instructed to type a fluent English output sentence into a text editor.

For each segment, the monolingual post-editor was instructed to record confidence according to the guidelines shown in in Table 3. Post-edited segments marked as "Verify" or "Partially unsure" were passed on to a bilingual post-editor to verify and correct, if necessary. Post-edited segments marked as "Completely unsure" were passed to a bilingual post-editor to post-edit or translate from scratch. Table 4 shows the breakdown of post-edited sentences by post-editor confidence. We observe that the monolingual domain expert post-edited 86.7% of the sentences without requesting assistance from a bilingual post-editor.

In determining confidence in a post-edited segment, we expect the monolingual post-editor to consider the segment's coherence with surrounding segments, and its semantic consistency with the entire document, taking into account the post-editor's own expertise in the domain. Because the monolingual post-editor does not know the source language, there is no guarantee that post-edited segments in which the post-editor is confident completely and correctly convey the meaning present in the respective source segments. For this reason, in Section 4 we perform a bilingual adequacy evaluation over all post-edited segments.

4 Evaluation

A high rate of post-editor confidence (as seen in Table 4) is worthy of note only if the post-editor's confidence is justified by corresponding high quality in post-edited results. Most machine translation experiments report quality according to BLEU or some other automated metric, as judged against one or more reference translations. In our case, the results of our work represent the only known translation of the document in question — as such, no reference trans-

lation is available.

4.1 Post-editor Confidence and Translation Adequacy

To determine the quality of post-edited translations, we asked a bilingual judge to rank the adequacy of the post-edited translations. The judge is a native English speaker fluent in French who was not involved in translating or post-editing any segments in this task. The bilingual judge was asked to rate the adequacy of all post-edited segments, using the evaluation guidelines shown in Table 5, which were adapted from Albrecht et al. (2009).

10	The meaning of the French sentence is fully conveyed in the English translation
8	Most of the meaning of the French sentence is conveyed in the English translation
6	The English translation misunderstands the French sentence in a major way, or has many small mistakes
4	Very little information from the French sentence is conveyed in the English translation
2	The English translation makes no sense at all

Table 5: Adequacy evaluation guidelines for bilingual human judges, adapted from Albrecht et al. (2009).

	Evaluation Category				
	2	4	6	8	10
# sentences	0	0	1	9	231
% sentences	0.0%	0.0%	0.4%	3.7%	95.9%

Table 6: Number and percentage of 241 evaluated sentences judged to be in each category by a bilingual judge. Category labels are defined in Table 5.

The resulting adequacy scores for all 241 post-edited segments are shown in Table 6. We observe that a very high percentage of post-edited segments (95.9% of segments) are rated by the bilingual judge to be completely correct translations of the original French. The remainder are either judged to be mostly correct (3.7% of segments) or partially correct (0.4% of segments). Of the 241 segments, the monolingual post-editor was confident in the post-editing of 209 segments. Those 209 segments were not shown to a bilingual post-editor; we observe that 96.5% of those 209 sentences, which were post-edited by only a monolingual post-editor, are judged to be completely correct by the bilingual judge.

Despite shortcomings (Callison-Burch et al., 2006), BLEU remains a widely used metric for MT evaluation. A somewhat conservative estimate on the quality of the post-edited translations can be measured using BLEU by treating the post-edited translations as a reference translation, and then treating as non-matches (for the purposes of calculating BLEU) all post-edited sentences whose bilingual adequacy score is less than 10; these results are shown in Table 7.

In addition, we cross-tabulate monolingual post-editor confidence (shown in Table 4) with bilingual adequacy judgments (shown in Table 6) to substantiate post-editor confidence with actual translation adequacy for each sentence. The results are shown in Table 8. These results

	BLEU	BLEU-cased
Post-edited (deleting non-perfect translations)	93.3	93.3

Table 7: Translation quality as measured by BLEU (Papineni et al., 2002) of the post-edited machine translation output, treating as non-matches (for the purposes of calculating BLEU) all post-edited sentences whose bilingual adequacy score is less than 10.

		Evaluation Category				
		2	4	6	8	10
Post-Editor Confidence	Completely unsure	0	0	0	0	8
	Partially Unsure	0	0	0	2	11
	Verify	0	0	0	1	10
	Confident	0	0	1	6	202

Table 8: For each of the 241 evaluated sentences, the adequacy category assigned by a bilingual judge, along with confidence assigned by the post-editor. Adequacy category labels are defined in Table 5. Confidence labels are defined in Table 3.

indicate that the high level of post-editor confidence is for the most part justified. Of the 209 segments where the post-editor was confident, only 7 were judged to be less than completely adequate translations.

For reference, these 7 segments are reproduced in their entirety in Appendix A. Four of the segments marked as less than completely adequate contain minor errors in typography or lexical choice. The post-edited translation of Segment 107 substitutes the more technical English term *the data* instead of the more literal *the work* or *the text* for the French term *l'oeuvre*. In segment 171, the English translates the French word *lecteurs* as *readers*, which is a valid translation for that French term, but is an incorrect lexical choice in context. Segment 196 incorrectly uses the literal translation *in the state* instead of a more appropriate idiomatic translation, such as *as is*, for the French phrase *en l'état*. Segment 215 consists entirely of a URL; the post-edited translation is rated 8 instead of 10, presumably because the English "translation" does not faithfully reproduce a spurious space character that appears in the French segment.

The remaining three segments contain more serious problems. The English translation of segment 8 elides a clause present in the French, resulting in an English translation that is perfectly fluent but semantically different from the original French. Segments 183 and 198 each contain phrases that are ill-formed in English, and also do not properly convey the semantic content of the respective French source segments.

4.2 Examining Machine Translation Results

Ideally, it would be desirable to evaluate the raw (un-edited) machine translation using the same 10-point adequacy metric used to evaluate the post-edited translations. Due to time constraints, we were unable to collect bilingual adequacy judgements on the raw (un-edited) machine translation output. We intend to pursue this in future work; in order to enable any interested researchers to perform such an analysis, we are making available for download both the post-edited results and the machine translation output of all three systems as supplementary materials to accompany this paper.

In the absence of a manual evaluation, we consider various automatic metrics in an attempt to provide at least some insight into the machine translation quality. Recall that over 95% of post-edited translations in our task were judged to be completely correct. Given this very high adequacy rate, we propose that it is not unreasonable to treat this post-edited data as reference

		Post-edited translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	52.6	51.8	17.70	35.23
	Systran	37.2	36.6	30.29	49.21
	Moses	14.0	11.8	67.34	87.09

Table 9: Similarity of the post-edited translations with the raw (un-edited) machine translation output from each MT system, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Google translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	100.0	100.0	0.0	0.0
	Systran	37.2	36.4	30.2	45.0
	Moses	17.6	15.1	65.6	80.2

Table 10: Similarity of the raw (un-edited) output of Google Translate with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Systran translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	37.3	36.6	27.9	41.5
	Systran	100.0	100.0	0.0	0.0
	Moses	21.0	17.9	56.0	72.6

Table 11: Similarity of the raw (un-edited) output of Systran with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

		Moses translations as reference			
		BLEU	BLEU-cased	PER	WER
System	Google Translate	17.4	14.9	51.2	62.6
	Systran	21.0	17.9	47.4	61.5
	Moses	100.0	100.0	0.0	0.0

Table 12: Similarity of the raw (un-edited) output of Moses with the raw (un-edited) machine translation output from the other two MT systems, as measured by case-insensitive and case-sensitive BLEU (Papineni et al., 2002), position-independent word error rate (PER), and word error rate (WER).

translations in order to examine the quality of the machine translation results used in this experiment. Treating the post-edited translations as a reference translation, we calculate BLEU, word error rate (WER), and position-independent word error rate (PER) on the output from the machine translation three systems.[2]

The results are shown in Table 9. Of the three MT systems, we observe the best scores for the raw MT output from Google Translate. Recall that the monolingual post-editor, when post-editing a segment, had the freedom to use the results of any of the three MT systems as the starting point for post-editing. The good automated metrics scores for Google Translate suggest that the post-editor drew most heavily from the Google Translate results when post-editing.

To get an indication of the relative similarity of the respective segments of the three MT systems, we also calculate BLEU, PER, and WER, treating (in turn) each MT system output as the reference for the purposes of automatic metric calculations. These results are shown in Tables 10, 11, and 12. We observe from these results that the output of Google Translate and Systran are somewhat similar, and that each of those systems differ substantially from the output of Moses.

5 Conclusion

The need for translation in today's highly connected and highly multilingual world far outstrips the supply of qualified human translators. In some cases of assimilation, where a user wants to extract information from a web page or other resource that is in a foreign language, imperfect machine translation can partially or completely satisfy the user's needs. In other more demanding cases of assimilation, as well as in most cases of dissemination, there is a need for a higher quality of translation than most machine translation systems provide.

Monolingual post-editing represents a middle ground between professional translation and raw use of machine translation. Previous work has indicated that monolingual post-editing can result in higher quality results than raw machine translation. In this work we have shown that when the monolingual post-editor is a domain expert in the material being translated, the monolingual post-editor can produce completely correct translations over 95% of the time. This work suggests that a monolingual post-editor can serve to effectively triage the translation process by forwarding on to bilingual post-editors only those segments which are too difficult for the monolingual post-editor to handle.

This work represents an initial examination into monolingual post-editing as a potential triage mechanism for translation. We plan a more thorough examination of this line of research. In future work, we plan to perform manual adequacy evaluations of the raw machine translation output in addition to the post-edited translations, in order to directly measure the adequacy improvements of monolingual post-editing. This work also is limited in scope by only making use of a single monolingual post-editor and a single document; future work should be broader in both of these dimensions, making use of multiple monolingual post-editors (both domain experts and non-experts) and multiple documents to be translated.

Acknowledgements

We wish to thank Katherine Young for her work post-editing, and Jeremy Gwinnup for his work training MT systems. Thanks also to Margaret Stanney and Patricia Phillips-Batoma for their help evaluating translations. Finally, substantial thanks to the anonymous reviewers. Your critiques and comments were extremely helpful, and have made this a better paper.

[2] We calculate BLEU using the `multi-bleu.pl` script from the Moses project, and calculate WER and PER using the `apertium-eval-translator.pl` script from the Apertium project (Forcada et al., 2011). Other metrics more directly tailored for post-editing scenarios, such as Joint Fuzzy Score (Zhechev, 2012) may also be useful to consider in future work.

Appendix

A Segments

- Segment 8 of 241 - Adequacy score 8

  French Les techniques actuelles de traduction automatique (TA) permettent de produire des traductions dont la qualité ne cesse de croitre.

  English Current machine translation (MT) techniques continue to improve.

- Segment 107 of 241 - Adequacy score 8

  French L'oeuvre, composée de 545 segments et 10731 mots est divisée en trois blocs identiques.

  English The data, made up of 545 segments and 10731 words was divided into three equal blocks.

- Segment 171 of 241 - Adequacy score 8

  French Après trois questions permettant de mieux cerner le profil du lecteur, une première partie (5 questions) interroge les lecteurs sur la lisibilité et la qualité du texte littéraire traduit.

  English After three questions to better understand the profile of the player, the first portion (5 questions) asks readers about readability and quality of the translated literary text.

- Segment 183 of 241 - Adequacy score 8

  French ce résultat mitigé indique peut-être un désintérêt de certains lecteurs pour les aspects les plus techniques de l'oeuvre.

  English this mixed result may indicate a lack of interest by some readers to the most technical of the work aspects.

- Segment 196 of 241 - Adequacy score 8

  French Le manque de place ne nous permet pas de commenter ces remarques mais nous pensons qu'elles sont assez explicites pour être délivrées en l'état.

  English Lack of space does not allow us to comment on these remarks but we think that they are sufficiently clear to be delivered in the state.

- Segment 198 of 241 - Adequacy score 6

  French Le texte auquel vous êtes parvenu restitue une image fidéle du contenu de l'article de Powers.

  English The text you have successfully reproduces faithfully the content of the article by Powers.

- Segment 215 of 241 - Adequacy score 8

  French `10. https ://fluidsurveys.com/surveys/manuela-cristina/un-livre-sur-moi-qualite-de-la-traduction/?TEST_DATA=`

  English `10. https://fluidsurveys.com/surveys/manuela-cristina/un-livre-sur-moi-qualite-de-la-traduction/?TEST_DATA=`

References

Albrecht, J. S., Hwa, R., and Marai, G. E. (2009). Correcting automatic translations through collaborations between MT and monolingual target-language users. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–68, Athens, Greece.

Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E., and Vilar, J.-M. (2009). Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Besacier, L. (2014). Traduction automatisée d'une oeuvre littéraire: une étude pilote. In *Actes de 21ème Traitement Automatique des Langues Naturelles (TALN '14)*, pages 389–394.

Callison-Burch, C. (2005). Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL 06)*, page 249256.

Denkowski, M. and Lavie, A. (2012). TransCenter: Web-based translation research suite. In *Proceedings of the AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*.

Forcada, M., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Google (2014). Google Translate. `http://translate.google.com`.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 439–448, Paris, France.

Kazi, M., Coury, M., Salesky, E., Ray, J., Shen, W., Gleason, T., Anderson, T., Erdmann, G., Schwartz, L., Ore, B., Slyh, R., Gwinnup, J., Young, K., and Hutt, M. (2013). The MIT-LL/AFRL IWSLT-2013 MT system. In *The 10th International Workshop on Spoken Language Translation (IWSLT'13)*, pages 136–143, Heidelberg, Germany.

Koehn, P. (2009a). A process study of computer aided translation. *Machine Translation*, 23(4):241–263.

Koehn, P. (2009b). A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20, Suntec, Singapore.

Koehn, P. (2010). Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545, Los Angeles, California.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07) Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Langlais, P., Foster, G., and Lapalme, G. (2000). TransType: A computer-aided translation typing system. In *Proceedings of the ANLP/NAACL 2000 Workshop on Embedded Machine Translation Systems*, pages 46–51, Seattle, Washington.

Mitchell, L., Roturier, J., and O'Brien, S. (2013). Community-based post-editing of machine translation content: monolingual vs. bilingual. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP-2)*, pages 35–43, Nice, France. EAMT.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL '02)*, pages 311–318, Philadelphia, Pennsylvania.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, 93:7–16.

Schwartz, L., Anderson, T., Gwinnup, J., and Young, K. M. (2014). Machine translation and monolingual postediting: The AFRL WMT-14 system. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 186–194, Baltimore, Maryland. Association for Computational Linguistics.

Systran (2010). Systran server 7.4.2. `http://www.systransoft.com`.

Zhechev, V. (2012). Machine Translation Infrastructure and Post-editing Performance at Autodesk. In *AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012)*, pages 87–96, San Diego, USA. Association for Machine Translation in the Americas (AMTA).